Paleobiology, 29(4), 2003, pp. 455–467

On the continuity of background and mass extinction

Steve C. Wang

Abstract.—Do mass extinctions grade continuously into the background extinctions occurring throughout the history of life, or are they a fundamentally distinct phenomenon that cannot be explained by processes responsible for background extinction? Various criteria have been proposed for addressing this question, including approaches based on physical mechanisms, ecological selectivity, and statistical characterizations of extinction intensities.

Here I propose a framework defining three types of continuity of mass and background extinctions—continuity of cause, continuity of effect, and continuity of magnitude. I test the third type of continuity with a statistical method based on kernel density estimation. Previous statistical approaches typically have examined quantitative characteristics of mass extinctions (such as metrics of extinction intensity) and compared them with the distribution of such characteristics associated with background extinctions. If mass extinctions are outliers, or are separated by a gap from background extinctions, the distinctness of mass extinctions is supported.

In this paper I apply Silverman's Critical Bandwidth Test to test for the continuity of mass extinctions by applying kernel density estimation and bootstrap modality testing. The method improves on existing work based on searching for gaps in histograms, in that it does not depend on arbitrary choices of parameters (such as bin widths for histograms), and provides a direct estimate of the significance of continuities or gaps in observed extinction intensities. I am thus able to test rigorously whether differences between mass extinctions and background extinctions are statistically significant.

I apply the methodology to Sepkoski's database of Phanerozoic marine genera. I conclude that mass and background extinctions appear to be continuous at this third level—continuity of magnitude—even though evidence suggests that they are discontinuous at the first and second levels.

Steve C. Wang. Department of Statistics, Harvard University, Cambridge, Massachusetts 02138
Present address: Department of Mathematics and Statistics, Swarthmore College, Swarthmore, Pennsylvania 19081. E-mail: scwang@swarthmore.edu

Accepted: 21 March 2003

Introduction

Ever since the proposal that the end-Cretaceous mass extinction resulted from a bolide impact (Alvarez et al. 1980), the nature of mass extinctions has been the subject of much debate. A key question is whether mass extinctions grade continuously into the background extinctions occurring throughout the history of life, or whether they constitute a fundamentally different phenomenon. In the former view, mass extinctions represent the right tail of a continuum, separated from background extinctions by an arbitrary cutoff, much as a blizzard and a flurry represent varying degrees in a continuum of snowfall. In the latter view, mass extinctions represent a distinct phenomenon that cannot be explained by merely scaling up background extinctions, just as a hailstorm is not merely a larger version of a snow flurry.

Both sides of the debate have their supporters. Quinn (1983), Raup (1986, 1991a,b, 1994), Bambach and Gilinsky (1986), and McKinney (1987) argued for the continuity of mass extinctions with background extinctions. On the other hand, Raup and Sepkoski (1982), Gould (1985), and Bambach and Knoll (2001) argued that mass extinctions are a distinct phenomenon, and Stigler (1987) also found evidence for this position. One reason for the debate stems from the fact that different authors have different meanings for what it means for mass extinctions to be ''continuous'' with or ''distinct'' from background extinctions, and their meanings are often stated only implicitly. Here I propose a framework defining three types of continuity—continuity of cause, continuity of effect, and continuity of magnitude. These three types of continuity are independent of each other, in that mass extinctions may be discontinuous at one level but continuous at the others. For the third type of continuity, I propose the application of a statisti-

 0094-8373/00/2904-0001/$1.00

cal method to test whether mass extinctions are continuous with background extinctions.

Types of Continuity

Continuity of cause occurs when the same processes that are responsible for background extinctions, operating at an increased level or intensity, also cause mass extinctions (Miller 1998). For example, if background extinctions are caused by terrestrial factors (e.g., changes in sea level and climate), and these same factors at a more extreme intensity also cause mass extinctions, then mass and background extinctions would be continuous in cause. Continuity of cause would also be established if mass extinctions result primarily from competition among clades, as Briggs (1998) argues. If, on the other hand, mass extinctions are caused by factors different from those causing background extinctions (e.g., bolide impact), this would constitute a discontinuity of cause. In the latter case, mass extinction would differ qualitatively from background extinction and in a fundamental way. The well-accepted evidence for an impact at the end of the Cretaceous (Alvarez et al. 1980) supports discontinuity of cause, as does the newer and more controversial evidence for impact at the ends of the Permian and Triassic Periods (Becker et al. 2001; Olsen et al. 2002). Raup and Boyajian's finding (1988) that major extinction events result from environmental disturbances may also be interpreted as supporting discontinuity at this level. Continuity of cause may be further subdivided into ultimate cause (e.g., bolide impact—the trigger mechanism) and immediate cause (e.g., resulting nutrient crisis—the kill mechanism).

Continuity of effect is established when background and mass extinctions exhibit common patterns of selectivity on taxonomic, functional, morphological, geographical, or other criteria. In other words, continuity of effect refers to whether the biological and ecological effects of background extinction and mass extinction are similar in nature, if not in magnitude. For example, McKinney (1987) found that extinction rates in background and mass extinctions are strongly correlated for ten major marine taxa, Erwin (1989, 1990) found no difference in selectivity of gastropods at the end-Permian mass extinction compared with background extinction patterns, and Boyajian (1991) found no differences in selectivity in mass extinctions with regard to taxon age after controlling for extinction size. On the other hand, several studies have found evidence for differential patterns of survival at mass extinctions compared with background extinction. These include studies by Anstey (1986) on Ordovician bryozoans, Jablonski (1986) on Cretaceous mollusks, Jablonski and Raup (1995) on Cretaceous bivalves, Johansen (1989) on Cretaceous brachiopods, and Westrop (1989) on Cambrian trilobites; see also Stanley 1987 for a general discussion. Such findings support a discontinuity of effect. Gould (1985) also supported discontinuity of effect, arguing that mass extinctions constitute a third ''tier'' distinct from and irreducible to within-species competition (the first tier) and species-level selection (the second tier).

Continuity of magnitude exists when the distribution of intensities of mass extinctions (as measured by the number of extinctions per unit time or some other metric) grade smoothly and continuously into the intensities of background extinctions. Quinn (1983), Bambach and Gilinsky (1986), McKinney (1987), and Raup (1986, 1991a,b, 1994) found that mass extinctions are continuous at this level, and Thackeray (1990) arrived at a similar conclusion for nine extinction events with the exception of the end-Cretaceous event. On the other hand, Raup and Sepkoski (1982) and Bambach and Knoll (2001) argued for discontinuity, with Stigler (1987) also finding evidence for the latter position.

A common method of determining continuity of magnitude is by examining histograms of extinction intensities of Phanerozoic stages. If mass extinctions are continuous in magnitude, then such a histogram should appear unimodal, with no apparent gaps separating mass extinctions in the right tail of the distribution from background extinctions (Raup 1986: Fig. 1, also cited in Jablonski 1989; Raup 1991b: Fig. 4–4; Raup 1994: Fig. 2, also cited in Jablonski 2001). On the other hand, if such a histogram appears bimodal, with mass extinctions forming a cluster of outliers in the right tail, we would infer that mass extinctions

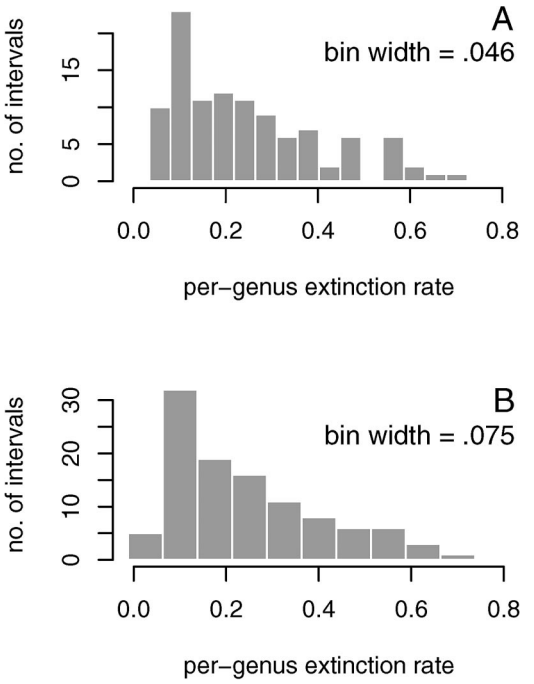

FIGURE 1. Histograms of per-genus proportional extinction rate for 107 Phanerozoic stages and substages from Sepkoski's compendium of marine genera, showing that the appearance of a histogram depends on the parameter values used in its construction. A, Using a bin width of 0.046 results in a gap between the largest extinctions (extinction rate ≤ 0.54) and other extinctions (extinction rate ≤ 0.49), with a peak around 0.55. B, When the same data are plotted with a bin width of 0.075, these features are not apparent.

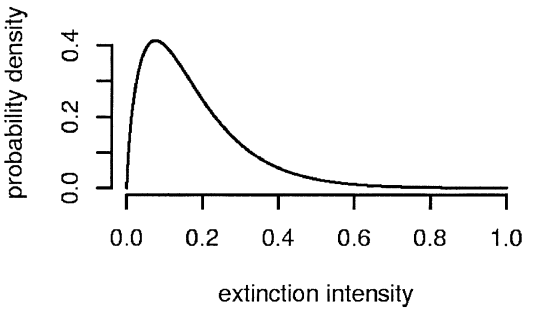

FIGURE 2. A hypothetical example of an extinction intensity curve $f(x)$. Such a curve is a probability density function (pdf) that models the underlying process governing extinction and describes how likely various intensities of extinctions are.

are a distinct phenomenon, discontinuous from background extinction (Raup and Sepkoski 1982; Bambach and Knoll 2001). In such analyses, an important question is how to rigorously determine if mass extinctions indeed constitute a second mode in a histogram of extinction intensities. This can be especially difficult because the appearance of a histogram—particularly the presence of modes, gaps, and outliers—depends on the arbitrary choice of parameters used to construct it, notably the bin width. Therefore, tests based on histograms can be unreliable, an issue I address in the next section.

Modes and Gaps in Histograms

In this section I discuss how a histogram's appearance depends on the choice of parameters. A histogram does not display exact numerical values, but rather places the data into ''bins.'' Raup (1994: Fig. 2) chose bins of {0–5%, 5–10%, . . . , 95–100%}. The choice of bin width and bin location is arbitrary: the bins {0–10%, 10–20%, . . . , 90–100%} would be as valid a selection, as would be {(-5)–5%, 5–15%, . . . , 95–105%}. Usually these choices are made automatically by software and have a relatively minor effect on the appearance of the resulting histogram. Some features of a histogram, however, are particularly sensitive to the choice of bin width and location parameters, notably the presence and location of modes, gaps, and outliers.

For instance, by using a large enough bin width—larger than the distance between any two neighboring points—a histogram will appear continuous (having no gaps between adjacent bars), even if outliers do exist. On the other hand, by using a small enough bin width, one can always create the appearance of a second mode, separated by a gap from the body of the data. As an example, consider Figure 1, which shows histograms of per-genus extinction rate for 107 Phanerozoic intervals (stages and substages) from Sepkoski's unpublished compendium of marine genera. Figure 1A uses a bin width of 0.046, whereas Figure 1B uses a bin width of 0.075. Although the data are identical in the two histograms, the distributions appear different, particularly in the right tail of the data. In Figure 1A, there is a second mode separated by a gap, with peaks at approximately 0.48 and 0.55 extinctions/genus, whereas in Figure 1B the data appear continuous with no additional peaks and no gap. Thus, in searching for evi-

dence of a second mode, we should use caution in drawing conclusions from histograms and other methods that are affected by the choice of arbitrary parameters.

Density Estimation

The histogram is just one way to display a distribution of extinction intensities. A histogram provides a discrete display—that is, a step function with jumps from each individual bar to the next. The true distribution of extinction intensities, however, is likely to be a smoothly varying curve, rather than a step function. In this section I describe how to construct a smooth estimate of the distribution of extinction intensities by using the statistical technique of *density estimation.*

Density estimation is a well-studied statistical technique (e.g., Silverman 1986). Here I give a brief conceptual introduction to the subject.

Suppose that the underlying process governing extinction intensities can be modeled by an intensity curve—a probability density function (pdf) describing how likely various intensities of extinctions are. Denote this intensity curve by $f(x)$, with x representing the intensity of an extinction. Such an extinction intensity curve might look something like the hypothetical curve in Figure 2. The goal is to estimate the underlying extinction intensity density curve $f(x)$, given a data set of observed extinction intensities.

I now describe how to create such a density estimate by contrasting it with the creation of a histogram. In constructing a histogram, two parameters must be specified: the width of each bin of the histogram and the location at which each bin is centered. (For simplicity, I assume bins are of equal widths.) Typically these parameters are set automatically by a software program and transparent to the user, but they must be explicitly chosen whether the user is aware of the choice or not. These parameters determine a fixed set of bins. The bars of a histogram graphically represent the number of observed data points that fall into each of these bins.

The process of constructing a simple density estimate is, in a sense, the opposite of that used to build a histogram. Instead of placing the data points into a fixed and predetermined set of bins, we place the bins according to the locations of the data points. Centered on each data point we place a bar, and then we sum the heights of these bars for all the data points (in a sense ''stacking'' the bars). The ''bar chart'' determined by this sum can be viewed as a simple density estimate, which I will denote as $\hat{f}_1(x)$. An example of this simple density estimate is shown for a data set of five points in Figure 3A.

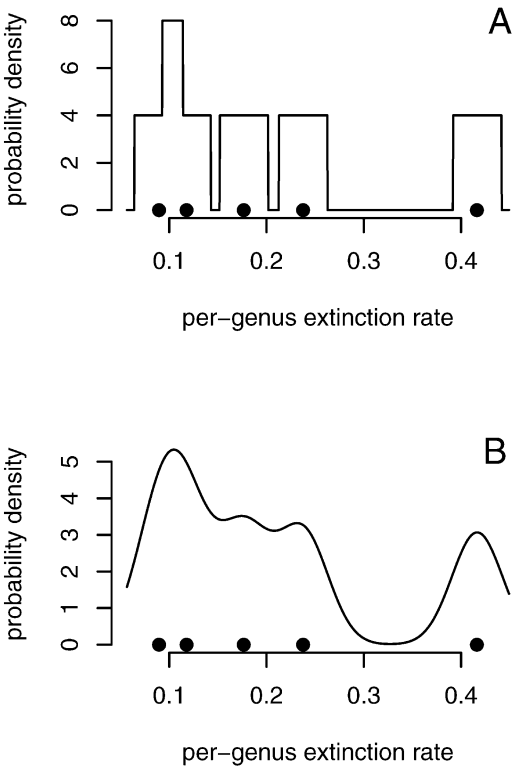

FIGURE 3. Density estimates for a random sample of five points. Five intervals were randomly chosen from the full data set of per-genus proportional extinction rates—middle Miocene (per-genus extinction rate = 0.09), Valanginian (0.12), Kimmeridgian (0.18), lower Atdabanian (0.24), upper Middle Cambrian (Upper) (0.42)—and are marked by black dots on the figures. A, Simple density estimate $\hat{f}_1(x)$ calculated with these five points. B, Kernel density estimate $\hat{f}(x)$ using a Normal kernel function. See text for details.

Formally, we can write this simple density estimate as follows. Let the n data points be denoted by $x_1, x_2, \ldots, x_n$. Define the function w_1 such that $w_1(y) = 1$ if $|y| \leq 1/2$, and $w_1(y) = 0$ otherwise. That is, w_1 represents a bar with height one, to be placed over each data point. For any point x on the real line, an es-

timate of the density function $f(x)$ at that point is given by summing over all bars according to the formula

$$\hat{f}_1(x) = \frac{1}{nh} \sum_{i=1}^{n} w_1\left(\frac{x - x_i}{h}\right). \quad (1)$$

Here, the parameter h is the bandwidth parameter, which plays a role equivalent to the bin width parameter in a histogram. The choice of h in a density estimate is arbitrary, as is the choice of bin width in a histogram, but I will show below that my conclusions about the continuity of mass extinctions do not depend on the choice of h. The factor $1/nh$ is a normalizing constant that ensures that $\hat{f}_1(x)$ integrates to one, as must be true of any pdf. Thus $\hat{f}_1(x)$ is a density estimate, an estimate of the probability density at the point x.

As I have defined it here, $\hat{f}_1(x)$ is not a particularly realistic estimate of the underlying extinction intensity curve. The true intensity curve $f(x)$ is most likely a smooth curve, not a step function as in Figure 3A: we expect the likelihood of various intensities of extinctions to grade smoothly, not jump abruptly from one intensity to the next. The ''blockiness'' of $\hat{f}_1(x)$ is a result of the choice of w_1, which is a step function. To avoid this blockiness, we can instead use a smooth function w integrating to one. Conceptually, this corresponds to placing a smooth curve on each data point and then summing up the curves to arrive at a density estimate.

Such a smooth function w satisfying $\int w(y)dy = 1$ is called a *kernel function*, and the resulting density estimate is called a *kernel density estimate*. Such a kernel density estimate $\hat{f}(x)$ is written as follows:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} w\left(\frac{x - x_i}{h}\right). \quad (2)$$

A common choice for the kernel function w is the Normal or Gaussian density function, $w(y) = 1/(2\pi)^{1/2}\exp(-y^2/2)$. Much research has been done in the statistics literature on the choice of kernel functions. The resulting density estimate is usually not overly sensitive to the particular kernel function chosen as long as certain conditions are met; often the Normal density is used because it has convenient mathematical properties. In the rest of the paper, all kernel density estimates will use the Normal kernel function. Figure 3B plots using the same data as in Figure 3A but with a Normal kernel function.

Further extensions to the kernel density estimate are possible. Using a fixed bandwidth h, for instance, sometimes results in too much smoothing near the center of the pdf but not enough smoothing in the tails. One approach to this problem is to adaptively vary the bandwidth, for example by letting the bandwidth equal the width spanned by the k nearest data points. I do not believe such a methodology is likely to affect my conclusions in this setting, so in this paper I will use kernel density estimates with fixed bandwidth.

Testing for Bimodality

If mass extinctions are a phenomenon qualitatively distinct from background extinctions, we would expect the underlying extinction intensity curve to have two modes or peaks separated by a gap. If mass extinctions grade smoothly into background extinctions, we would expect the intensity curve to have a long right tail but only one mode and no gaps. Of course, we do not know the form of the actual underlying intensity curve. However, we can estimate it by using a kernel density estimate, and then see whether the resulting kernel density estimate has one mode or two.

The appearance of a density estimate—including how many modes it has—depends on the bandwidth parameter h, analogous to the bin width of a histogram. With a large enough h the density estimate can always be made to appear unimodal; with a small enough h the density estimate can always be made to appear bimodal (or even multimodal, having more than two peaks). Therefore, the number of modes is inversely related to the size of the bandwidth h. (This property is not true for all kernel functions, but it is true when the Normal pdf is used as the kernel function [Silverman 1981].)

Given a particular data set, suppose we construct a series of density estimates starting with a small value of h and increasing to a large value of h. These density estimates will initially be bimodal or multimodal, but will become unimodal once h increases past a cer-

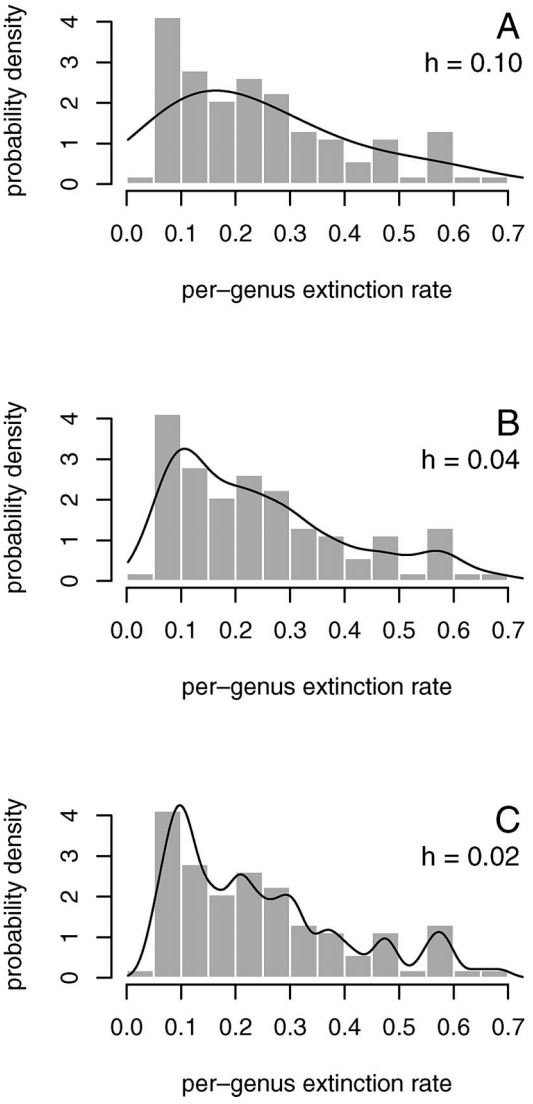

FIGURE 4. Histograms of per-genus proportional extinction rate and kernel density estimates. As the bandwidth h used in constructing the density estimate $\hat{f}(x)$ decreases, the number of modes of $\hat{f}(x)$ increases. The critical bandwidth h_{crit} is the smallest value of h for which the density estimate $\hat{f}(x)$ appears unimodal, and in this case would lie between 0.10 and 0.04. A, With $h = 0.10$, $\hat{f}(x)$ is unimodal. B, With $h = 0.04$, $\hat{f}(x)$ is bimodal (second mode appears near 0.58). C, With $h = 0.02$, $\hat{f}(x)$ is multimodal.

tain critical value (Fig. 4). Define h_{crit} as the value of h at which the density estimates change from bimodal to unimodal. That is, for $h > h_{crit}$, the density estimate will be unimodal, and for $h < h_{crit}$, the density estimate will be bimodal (or multimodal). Here h_{crit} is called the critical bandwidth. This idea was introduced by Silverman and is the basis for his Critical Bandwidth Test for bimodality (Silverman 1981, 1986; see also Efron and Tibshirani 1993), which I now describe.

If the underlying extinction intensity curve is truly bimodal, we can still construct a density estimate that appears unimodal, but we will need a very large bandwidth h because a high degree of smoothing will be necessary to make the density estimate appear unimodal. In other words, h_{crit} will be large if the intensity curve is truly bimodal. On the other hand, if the extinction intensity curve is truly unimodal, a density estimate will appear unimodal even with small values of h, because little smoothing will be necessary to make the density estimate appear unimodal. In other words, h_{crit} will be small if the intensity curve is truly unimodal.

We can therefore infer the true modality of the extinction intensity curve from the size of h_{crit}. That is, h_{crit} can serve as the test statistic in a hypothesis test of the modality of $f(x)$. The null hypothesis is that the true extinction intensity curve $f(x)$ is unimodal; the alternative hypothesis is that $f(x)$ is multimodal. Large values of h_{crit} provide evidence against the null hypothesis.

For a given data set, the value of h_{crit} can be calculated by using a binary search method on a computer. A natural question is, how large must h_{crit} be to reject the null hypothesis, providing statistically significant evidence of a multimodal extinction intensity curve? Equivalently, we might ask for the p-value corresponding to an observed value of h_{crit}. Critical values for statistical significance and p-values can be approximated by a bootstrap-based simulation; see the Appendix for details.

Results

I applied the Critical Bandwidth Test to data from J. J. Sepkoski's unpublished compendium of Phanerozoic marine genera (kindly provided by R. Bambach). All genera are included, with subgenera also included for mollusks. The number of extinctions and originations and total diversity are given for each of 107 intervals, at the stage and substage level, from the Nemakit–Daldynian (Lower Cambrian) to the Pliocene (Tertiary). All analyses

TABLE 1. Extinction metrics. Note that estimated standing diversity = total diversity − 1/2 originations − 1/2 extinctions.

Extinction intensity metric	Definition
A. Number of extinctions/interval	Extinctions
B. Number of extinctions/Myr	extinctions/interval duration
C. Per-genus proportional extinctions/interval	extinctions/total diversity
D. Per-genus proportional extinctions/Myr	(extinctions/total diversity)/interval duration
E. Van Valen metric/interval	extinctions/estimated standing diversity
F. Van Valen metric/Myr	(extinctions/estimated standing diversity)/interval duration

were run with the software R (Version 1.6.0) and Data Desk (Version 6.1) on an Apple Macintosh G4.

I used six common metrics for measuring extinction, which are defined in Table 1. All of these metrics measure extinction intensity, although they differ in whether or not they attempt to normalize for time, total diversity, standing diversity, or some combination thereof. Foote (1994) discussed these extinction intensity metrics, using simulations to explore their properties under various scenarios. He does not recommend any single metric, but rather finds that each has strengths and weaknesses, and each may give biased estimates of the true per-genus likelihood of extinction under certain conditions (for instance, when stratigraphic stage boundaries are defined by extinctions, as is the case for the standard geologic timescale). He does recommend against the use of proportional extinction per million years (metric D), which is biased under many realistic scenarios. See also the discussion in Raup 1986 and Raup and Boyajian 1988.

Histograms of the data using each of the six metrics are shown in Figure 5. Each plot also shows the kernel density estimate $\hat{f}(x)$ with bandwidth $h = h_{\text{crit}}$. With five of the six metrics, I was unable to reject the null hypothesis of unimodality. Only for metric A (number of extinctions per interval) was the discontinuity significant ($p = 0.01$). However, this metric is simply a raw count of the number of genera going extinct in the interval, without accounting for total diversity. The significance of the result found using this metric is due to one outlier, the Maastrichtian, in which nearly 1500 genera went extinct. Because total diversity was higher in the late Mesozoic by a factor of two or three compared with most times earlier in the Phanerozoic, it is not surprising that a large number of genera would be affected by a catastrophic event such as a bolide impact at this time. In fact, the number of genera lost in the Maastrichtian exceeded the total diversity of all but 21 of the previous 94 intervals in the database. Thus, using the raw number of extinctions per interval is misleading, and the significance of this result should carry little weight. The second-lowest p-value was for the number of extinctions per million years (metric B), with $p = 0.17$. This metric is merely a time-normalized version of metric A and therefore suffers from the same problems. The p-values for metrics C–F were 0.54, 0.48, 0.99, and 0.48, respectively; none were close to attaining statistical significance.

It has been noted (Raup and Sepkoski 1982; Thackeray 1990; Gilinsky 1994; Newman and Eble 1999) that extinction intensity has declined over the Phanerozoic. This would make any methodology overly conservative: a more recent extinction might be separated by a gap from recent background extinction intensities, but not from older background intensities. To control for this effect, I reran the above analyses with time-adjusted (detrended) intensities for all six metrics, using residuals from a linear regression of intensity on time. The results did not differ substantially. The p-values for metrics A–F were 0.04, 0.31, 0.48, 0.70, 0.81, and 0.24, respectively (Figure 6).

(I also tried to control for the decline in extinction intensities by rerunning the above analyses using residuals from a nonlinear lowess fit of intensity on time. However, these residuals included some negative outliers in the left tail of the distribution—that is, intervals with unusually *low* extinction rates. In some cases, these outliers formed a second

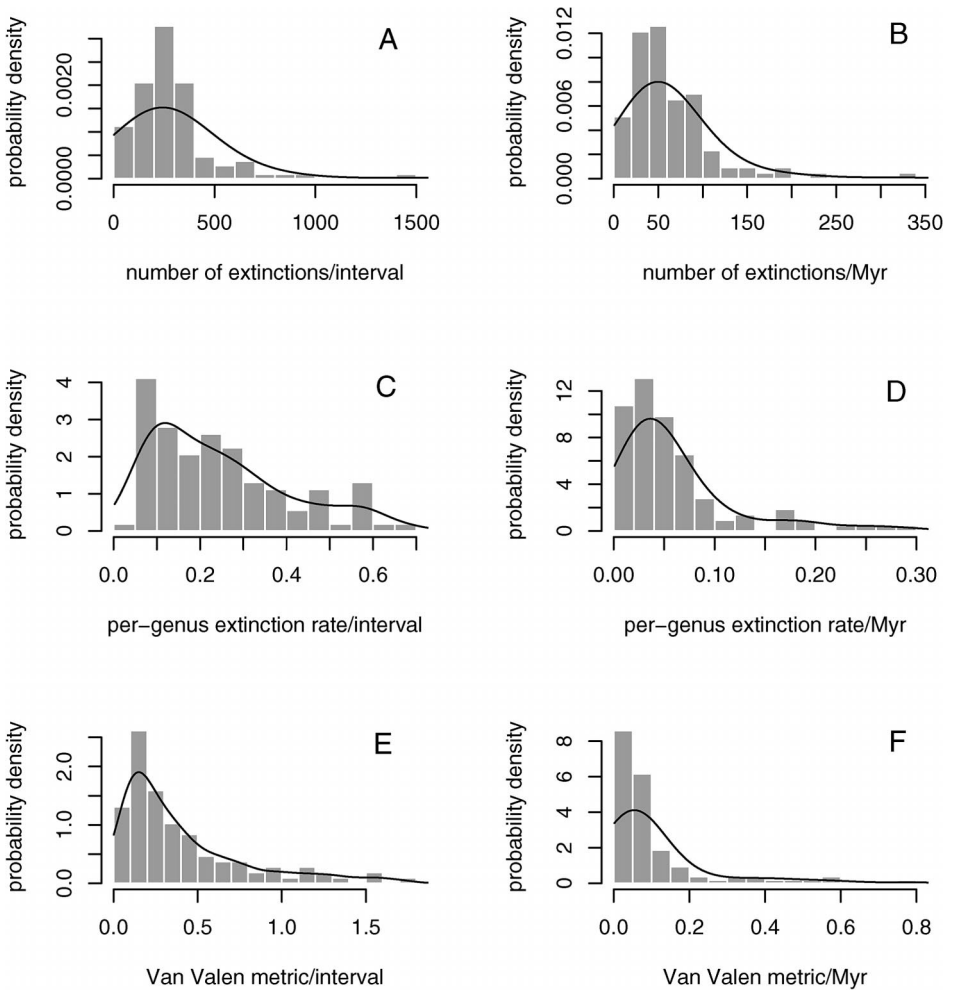

FIGURE 5. Histograms of extinction intensity metrics, with kernel density estimates of the underlying extinction intensity curve. Kernel density estimates are constructed with bandwidth equal to h_{crit}. A, Number of extinctions per interval. The p-value for testing the null hypothesis that the underlying extinction intensity curve is unimodal, indicating continuity of magnitude, was $p = 0.01$. B, Number of extinctions per million years ($p = 0.17$). C, Per-genus proportional extinctions per interval ($p = 0.54$). D, Per-genus proportional extinctions per million years ($p = 0.48$). E, Van Valen metric per interval ($p = 0.99$). F, Van Valen metric per million years ($p = 0.48$). The three highest intervals for each metric are as follows: A, Maastrichtian, upper Ashgillian, Guadalupian. B, Upper Ashgillian, Maastrichtian, lower Botomian. C, Djulfian, Dresbachian, upper Tommotian. D, Upper Tommotian, lower Botomian, upper Botomian. E, Dresbachian, upper Tommotian, Franconian. F, Upper Tommotian, lower Botomian, upper Atdabanian.

mode that was sufficient to cause the Critical Bandwidth Test to reject the null hypothesis of unimodality. In such cases, the Critical Bandwidth Test is not appropriate for testing the continuity of *mass* extinctions, because the bimodality indicated the presence of intervals with greatly reduced rather than elevated extinction intensities.)

The weight of the evidence suggests that observed gaps and outliers in these histograms are not statistically significant. Note, however, that although a significant result would certainly suggest a discontinuity of cause, a nonsignificant result may not necessarily suggest a continuity of cause. For instance, it may be possible that mass and background extinctions are due to different causes, but the variation resulting from each cause may be so large that two distinct modes are not formed. To the extent that any conclusions can be drawn, these findings do support the position that mass extinctions are the right tail of a

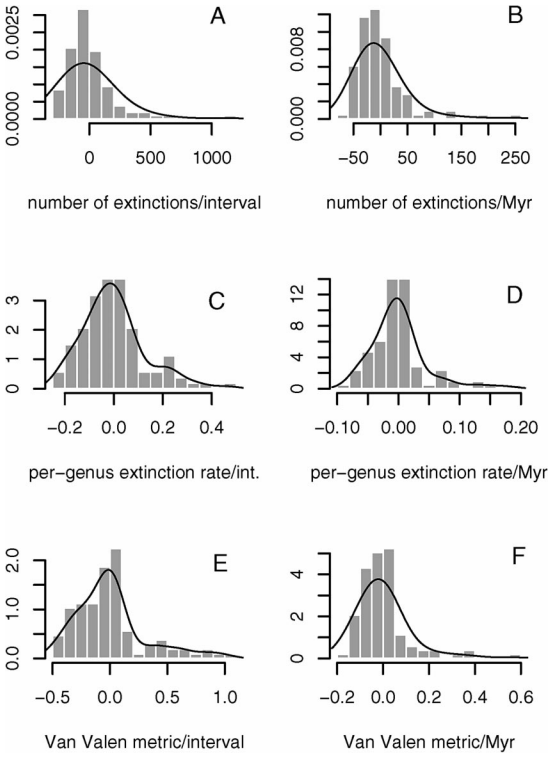

FIGURE 6. Histograms of time-adjusted extinction intensity metrics, with kernel density estimates of the underlying extinction intensity curve. Kernel density estimates are constructed with bandwidth equal to h_{crit}. To account for the Phanerozoic decline in extinction intensities, the data used here are residuals from a linear regression of intensity on time. Vertical axis indicates probability density; label has been omitted to save space. A, Number of extinctions per interval. The p-value for testing the null hypothesis that the underlying extinction intensity curve is unimodal, indicating continuity of magnitude, was $p = 0.04$. B, Number of extinctions per million years ($p = 0.31$). C, Per-genus proportional extinctions per interval ($p = 0.48$). D, Per-genus proportional extinctions per million years ($p = 0.70$). E, Van Valen metric per interval ($p = 0.81$). F, Van Valen metric per million years ($p = 0.24$). The three highest intervals for each metric are as follows: A, Maastrichtian, upper Ashgillian, Guadalupian. B, Upper Ashgillian, Maastrichtian, upper Eocene. C, Djulfian, Maastrichtian, Guadalupian. D, Upper Tommotian, lower Botomian, upper Botomian. E, Dresbachian, Djulfian, upper Tommotian. F, Upper Tommotian, lower Botomian, upper Atdabanian.

spectrum of continuous extinction intensities, and are not qualitatively different in magnitude from background extinctions. I thus find evidence to support a continuity of magnitude between mass extinctions and background extinctions at the timescale resolution of the Sepkoski database.

The Power of the Test

Failure to reject the null hypothesis does not necessarily imply acceptance of the null hypothesis. Instead, failure to reject the null hypothesis may also result from a lack of statistical power—the ability of a test to reject the null hypothesis when the alternative hypothesis is in fact true. To evaluate the power of the Critical Bandwidth Test, I carried out a series of simulations.

To calculate the power of a hypothesis test, a particular alternative hypothesis must be specified. That is, we must specify the shape of the true extinction intensity curve $f(x)$, including the locations and relative sizes of its modes. Of course, $f(x)$ is unknown, but we can construct several plausible hypothesized candidates for $f(x)$ and calculate the power of the test under each one. Figure 7 shows three such hypothesized extinction intensity curves for per-genus proportional extinctions per interval. (I have constructed scenarios for this metric only; results for other metrics would be equivalent because the simulation depends only on the shape of the curve and not the nature of the metric used.) The hypothesized extinction intensity curves in Figure 7A–C show progressively less differentiation between the modes representing background and mass extinction.

The hypothesized extinction intensity curve in Figure 7A was constructed assuming that 14 of the 107 intervals represent mass extinctions, and the other 93 background extinctions. Although this may sound like a large number of mass extinctions, note that the interval with the 14th-highest per-genus extinction rate was the Maastrichtian, and the 15th-highest the upper Norian (the uppermost interval of the Triassic in this timescale), both usually considered among the ''Big Five'' mass extinctions. I simulated 100 data sets under this scenario, each data set consisting of 107 per-genus extinction rates (representing the 107 Phanerozoic intervals) sampled from the extinction intensity curve shown in Figure 7A. I then applied the Critical Bandwidth Test to these 100 simulated data sets. With a significance level of $\alpha = 0.05$, the null hypothesis of unimodality was correctly rejected 99% of

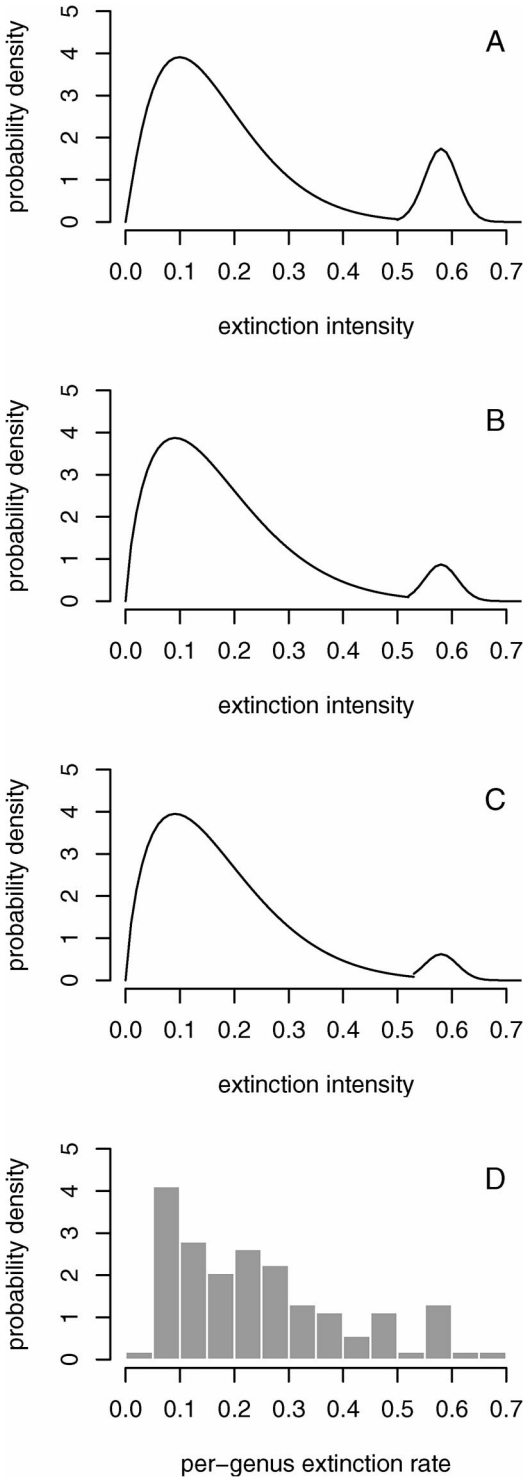

FIGURE 7. Hypothesized extinction intensity curves used in the power analysis for the Critical Bandwidth Test. The metric used is per-genus proportional extinctions per interval; results depend only on the shape of the curve and will hold for any metric. A, Hypothesized curve constructed assuming that mass extinctions are represented by the 14 highest of the 107 intervals, and background extinctions by the other 93 intervals. In 100 data sets simulated from this hypothesized curve, the null hypothesis of unimodality was correctly rejected 99% of the time when using a significance level of $\alpha = 0.05$, and 100% of the time when $\alpha = 0.10$. B, Hypothesized curve constructed assuming that mass extinctions are represented by the seven highest intervals, and background extinctions by the other 100 intervals. In 100 data sets simulated from this hypothesized curve, the null hypothesis of unimodality was correctly rejected 67% of the time when $\alpha = 0.05$ and 83% of the time when $\alpha = 0.10$. C, Hypothesized curve constructed assuming that mass extinctions are represented by the five highest intervals, and background extinctions by the other 102 intervals. In 100 data sets simulated from this hypothesized curve, the null hypothesis of unimodality was correctly rejected 30% of the time when $\alpha = 0.05$ and 46% of the time when $\alpha = 0.10$. D, Histogram of actual per-genus extinction rates, for comparison with the hypothesized extinction intensity curves in A–C.

the time. When a less strict significance level of $\alpha = 0.10$ was used, the null hypothesis of unimodality was correctly rejected 100% of the time. Therefore, if the true extinction intensity curve $f(x)$ is similar to the curve in Figure 7A, the test will have very high power.

The hypothesized extinction intensity curve in Figure 7B was constructed assuming that seven of the 107 intervals represent mass extinctions. (The actual intervals with the seven highest per-genus extinction rates are the Djulfian, Dresbachian, upper Tommotian, Franconian, Trempealeauan, lower upper Middle Cambrian, and upper Ashgillian.) I chose seven mass extinction intervals so that the resulting hypothesized curve in Figure 7B would have a second mode resembling the second mode in the histogram of actual per-genus extinction rates (shown in Figure 7D for comparison). In 100 data sets simulated under this scenario, the null hypothesis of unimodality was correctly rejected 67% of the time when a significance level of $\alpha = 0.05$ was used, and 83% of the time when $\alpha = 0.10$. Here again, if the true extinction intensity curve $f(x)$ is similar to the curve in Figure 7B, the test will have fairly high power.

The hypothesized extinction intensity curve in Figure 7C was constructed assuming that five of the 107 intervals represent mass extinctions. (The actual intervals with the five highest per-genus extinction rates were listed in

the previous paragraph.) In 100 data sets simulated under this scenario, the null hypothesis of unimodality was correctly rejected 30% of the time when a significance level of $\alpha = 0.05$ was used, and 46% of the time when $\alpha = 0.10$. In this case, then, the power of the test is lower.

Clearly, the power of the test is very sensitive to the shape of the true extinction intensity curve $f(x)$. If $f(x)$ resembles the hypothesized curve in Figure 7C, the test will have low power and may not reject the null hypothesis (with a sample size of 107 intervals) even though $f(x)$ is truly bimodal. If $f(x)$ has slightly stronger bimodality, resembling the hypothesized curve in Figure 7B, the test will have much higher power and is likely to correctly reject the null hypothesis. If $f(x)$ has bimodality as pronounced as that of the hypothesized curve in Figure 7A, the test is virtually certain to correctly reject the null hypothesis. Which choice of $f(x)$ is most realistic? Obviously, we cannot know what the real $f(x)$ looks like. I believe that the curve in Figure 7B, with seven intervals representing mass extinctions, is a reasonable guess. In that case, because the statistical power of the test is then high, we can be confident that the results observed here reflect a true unimodality of the extinction intensity curve.

Discussion and Conclusions

I applied Silverman's Critical Bandwidth Test to various extinction metrics. This methodology provides a direct statistical test of significance of apparent modes and gaps in a distribution of extinction intensities, thus allowing us to rigorously test the continuity of mass extinctions and background extinctions. The test does not depend on arbitrary choices of parameters, as is true of other methodologies (e.g., searching for gaps in histograms).

As an aside, Silverman's Critical Bandwidth Test may be useful in other contexts in paleobiology. The test can be applied to any situation in which one wants to determine if an observed distribution is a mixture of more than one subgroup. For instance, the test can be used to determine whether two distinct species are present in a collection of specimens (e.g., Webster 2001).

The results of my analysis (and any similar analysis) depend strongly on the temporal resolution of the data. The data used here are resolved to the stage or substage level (approximately 2–10-Myr bins for most intervals); therefore, the results suggest that extinctions are continuous in magnitude at substage-level resolution. Mass extinctions may occur on much shorter timescales, in which case Sepkoski's data combine times of mass extinction and times of background extinction within a single interval. It is possible that Phanerozoic extinctions are discontinuous in magnitude at a finer level of resolution, a possibility not at odds with my results here. If a future data set is compiled with finer temporal resolution, this analysis could be repeated with that data set as well.

In addition to the Critical Bandwidth Test, other methods can be used to test whether background and mass extinction are continuous in magnitude. Stigler (1987) uses a likelihood-based method. Similar approaches, using a likelihood criterion to evaluate mixture models, have been used by Hunt and Chapman (2001) to detect instar clusters in arthropod size distributions, and by Monchot and Léchelle (2002) to detect sexual dimorphism in Pleistocene bovines. Such approaches could be applied here as well, and in fact Hunt and Chapman (2001) suggested distinguishing background and mass extinctions as another application of their methodology. A potential disadvantage of likelihood-based methods is that they require the specification of an explicit statistical model for the distribution of the characteristic under study. For instance, both Hunt and Chapman (2001) and Monchot and Léchelle (2002) assumed that sizes within each subgroup (each instar or each sex) are normally distributed. An assumption of normality would be untenable in our case, however, because the distribution of background extinction intensities is strongly skewed (another parametric family such as the Gamma distribution may be more plausible). In contrast, the Critical Bandwidth Test has the advantage of being nonparametric, so that one not need make any assumptions about the distributional form of the data. On the other hand, likelihood-based methods may have higher power because of their parametric as-

sumptions. Further comparison of these approaches in paleontological contexts is a topic that merits further study.

With the consensus favoring the hypothesis of bolide impact as a cause of the end-Cretaceous mass extinction, and more tenuous recent evidence that other mass extinctions may have been caused by bolide impact (Becker et al. 2001; Olsen et al. 2002), it seems clear that at least some mass extinctions are discontinuous in cause from background extinctions. Evidence also exists for a discontinuity of effect, with studies (Anstey 1986; Jablonski 1986; Jablonski and Raup 1995; Johansen 1989; Westrop 1989) finding differing patterns of selectivity between mass extinctions and background extinctions. Nonetheless, despite these discontinuities in cause and effect, my findings here support a continuity of magnitude. It is intriguing that such a discontinuity of cause can produce a continuity of magnitude—that the effect of a bolide impact, though more extreme, nonetheless produces results that are not qualitatively different from results produced by terrestrial factors.

Acknowledgments

I thank S. Chang, R. Bambach, M. Benton, A. Bush, M. Foote, A. Knoll, and J. Payne for their helpful comments on drafts of this manuscript, and R. Bambach, M. Foote, and D. Jablonski for providing data. I am also grateful to J. Alroy, M. Dinmore, M. Hutcheson, C. Marshall, S. Porter, D. Raup, and S. Stigler for their assistance, and to the Swarthmore College Research Fund and the Lindback Foundation for financial support.

Literature Cited

Alvarez, L. W., W. Alvarez, F. Asaro, and H. V. Michel. 1980. Extraterrestrial cause for the Cretaceous-Tertiary extinction. Science 208:1095–1108.

Anstey, R. L. 1986. Bryozoan provinces and patterns of generic evolution and extinction in the late Ordovician of North America. Lethaia 19:33–51.

Bambach, R. K., and N. L. Gilinsky. 1986. Perspectives on the distribution of origination and extinction during the Phanerozoic. Geological Society of America Abstracts with Programs 18:534.

Bambach, R. K., and A. H. Knoll. 2001. Is there a separate class of ''mass'' extinctions? Geological Society of America Abstracts with Programs 33:A-141.

Becker, L., R. J. Poreda, A. G. Hunt, T. E. Bunch, and M. Rampino. 2001. Impact event at the Permian-Triassic boundary: evidence from extraterrestrial noble gases in fullerenes. Science 291:1530–1533.

Boyajian, G. E. 1991. Taxon age and selectivity of extinction. Paleobiology 17:49–57.

Briggs, J. C. 1998. Biotic replacements: extinction or clade interaction? BioScience 48:389–395.

Efron, B. W., and R. J. Tibshirani. 1993. An introduction to the bootstrap. Chapman and Hall, London.

Erwin, D. H. 1989. Regional paleoecology of Permian gastropod genera, southwestern United States and the end-Permian mass extinction. Palaios 4:424–38.

———. 1990. Carboniferous-Triassic gastropod diversity patterns and the Permo-Triassic mass extinction. Paleobiology 16:187–203.

Foote, M. 1994. Temporal variation in extinction risk and temporal scaling of extinction metrics. Paleobiology 20:424–444.

Gilinsky, N. L. 1994. Volatility and the Phanerozoic decline of background extinction intensity. Paleobiology 20:445–458.

Gould, S. J. 1985. The paradox of the first tier: an agenda for paleobiology. Paleobiology 11:2–12.

Hunt, G., and R. E. Chapman. 2001. Evaluating hypotheses of instar-grouping in arthropods: a maximum likelihood approach. Paleobiology 27:466–484.

Jablonski, D. 1986. Background and mass extinctions: the alternation of macroevolutionary regimes. Science 231:129–133.

———. 1989. The biology of mass extinction: a paleontological view. Philosophical Transactions of the Royal Society of London B 325:357–368.

———. 2001. Lessons from the past: evolutionary impacts of mass extinctions. Proceedings of the National Academy of Sciences USA 98:5393–5398.

Jablonski, D., and D. M. Raup. 1995. Selectivity of end-Cretaceous marine bivalve extinctions. Science 268:389–391.

Johansen, M. B. 1989. Background extinction and mass extinction of the brachiopods from the chalk of northwest Europe. Palaios 4:243–250.

McKinney, M. L. 1987. Taxonomic selectivity and continuous variation in mass and background extinctions of marine taxa. Nature 325:143–145.

Miller, A. I. 1998. Biotic transitions in global marine diversity. Science 281:1157–1160.

Monchot, H., and J. Léchelle. 2002. Statistical nonparametric methods for the study of fossil populations. Paleobiology 28: 55–69.

Newman, M. E. J., and G. J. Eble. 1999. Decline in extinction rates and scale invariance in the fossil record. Paleobiology 25:434–439.

Olsen, P. E., D. V. Kent, H.-D. Sues, C. Koeberl, H. Huber, A. Montanari, E. C. Rainforth, S. J. Fowell, M. J. Szajna, and B. W. Hartline. 2002. Ascent of dinosaurs linked to an Iridium anomaly at the Triassic-Jurassic boundary. Science 296:1305–1307.

Quinn, J. F. 1983. Mass extinctions in the fossil record. Science 219:1239–1240.

Raup, D. M. 1986. Biological extinction in earth history. Science 231:1528–1533.

———. 1991a. A kill curve for Phanerozoic marine species. Paleobiology 17:37–48.

———. 1991b. Extinction: bad genes or bad luck? Norton, New York.

———. 1994. The role of extinction in evolution. Proceedings of the National Academy of Sciences USA 91:6758–6763.

Raup, D. M., and G. E. Boyajian. 1988. Patterns of generic extinction in the fossil record. Paleobiology 14:109–125.

Raup, D. M., and J. J. Sepkoski Jr. 1982. Mass extinctions in the marine fossil record. Science 215:1501–1503.

Silverman, B. W. 1981. Using kernel density estimates to inves-

tigate multimodality. Journal of the Royal Statistical Society B 43:97–99.

———. 1986. Density estimation for statistics and data analysis. Chapman and Hall, New York.

Stanley, S. M. 1987. Extinction. Scientific American Books, New York.

Stigler, S. M. 1987. Testing hypotheses or fitting models? Another look at mass extinctions. Pp. 147–159 *in* M. H. Nitecki and A. Hoffman, eds. Neutral models in biology. Oxford University Press, New York.

Thackeray, J. F. 1990. Rates of extinction in marine invertebrates: further comparison between background and mass extinction. Paleobiology 16:22–24.

Webster, M. 2001. Intraspecific variation and morphological evolution in the early Cambrian trilobite Bristolia (Olenelloidea). Geological Society of America Abstracts with Programs 33:A-31.

Westrop, S. R. 1989. Macroevolutionary implication of mass extinction: evidence from an upper Cambrian stage boundary. Paleobiology 15:46–52.

Appendix

In this section I briefly describe how to calculate *p*-values and critical values for the statistical significance of the random variable h_{crit} in Silverman's Critical Bandwidth test. For details, see references by Silverman (1981, 1986) and Efron and Tibshirani (1993).

The goal is to determine how large h_{crit} must be to provide statistically significant evidence of a multimodal extinction intensity curve. Alternatively, we may want to determine a *p*-value representing the significance of our observed value of h_{crit}. By definition, the *p*-value is the probability of exceeding the observed value of h_{crit} if the null hypothesis were true. (Recall that the null hypothesis is that the extinction intensity curve $f(x)$ is unimodal.) To calculate the *p*-value, we need to determine the sampling distribution of the random variable h_{crit} under the null hypothesis. In this case, however, because h_{crit} is not a simple function of $f(x)$, this problem is mathematically intractable and the sampling distribution cannot be found analytically.

I instead approximate the sampling distribution of h_{crit} under the null hypothesis by using a bootstrap-based simulation. Here another difficulty arises. Our null hypothesis (that $f(x)$ is unimodal) is a broad one, because the number of unimodal distributions is infinite. Further, it may not plausible to assume that $f(x)$ belongs to a particular parametric family (e.g., Normal). Thus we cannot simply simulate data sets sampled from $f(x)$ under the null hypothesis, because $f(x)$ is not uniquely specified.

Instead of simulating data sets sampled from $f(x)$, I will simulate data sets sampled from the density estimate $\hat{f}(x)$ that best represents $f(x)$ under the null hypothesis. What does such an $\hat{f}(x)$ look like? Clearly $\hat{f}(x)$ must be unimodal, and it should be as ''close'' as possible to the actual data. To satisfy these conditions, I use the $\hat{f}(x)$ that is constructed from the observed data using bandwidth $h = h_{crit}$. That is, I simulate data sets sampled from an estimate of $f(x)$ that is unimodal but closest to being bimodal, and based on the observed data. Then, for each simulated data set, I calculate and save the value of h_{crit} from that data set. With enough simulated data sets, I am able to estimate the sampling distribution of h_{crit} under the null hypothesis.

I can then determine the *p*-value associated with any observed value of h_{crit} by checking how often the simulated values of h_{crit} exceed that observed value. Similarly, to calculate the critical value of h_{crit} needed to achieve statistical significance at, say, the $\alpha = 0.05$ level, I can find the value of h_{crit} that is exceeded by only the largest 5% of the simulated values. I thus arrive at a criterion for determining if a particular value of h_{crit} is ''large'': if such a value would occur by chance less than 5% of the time when sampling from the unimodal $\hat{f}(x)$, then the null hypothesis of unimodality is rejected.